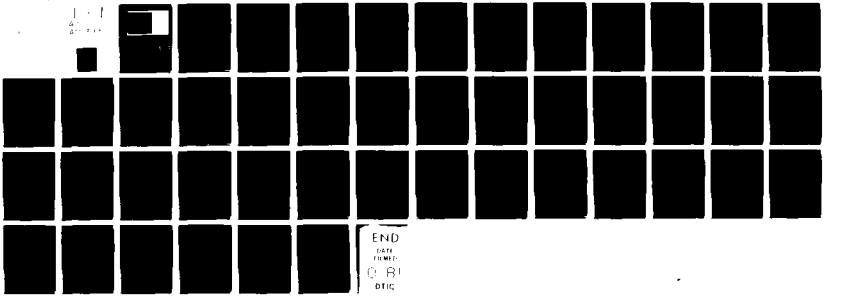
AD-A103 875 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
ON THE ESTIMATION OF A PROBABILITY DENSITY FUNCTION BY THE MAXI--ETC(U)
JUN 81 B W SILVERMAN DAAG29-80-C-0041
UNCLASSIFIED MRC-TSR-2228 NL



END
DATE FILMED
9 81
DTIC

MRC Technical Summary Report #2228

ON THE ESTIMATION OF A PROBABILITY DENSITY FUNCTION BY THE MAXIMUM PENALIZED LIKELIHOOD METHOD

B. W. Silverman

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

June 1981

(Received March 25, 1981)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

ON THE ESTIMATION OF A PROBABILITY DENSITY FUNCTION BY THE MAXIMUM PENALIZED LIKELIHOOD METHOD

B. W. Silverman*

Technical Summary Report #2228
June 1981

ABSTRACT

A class of probability density estimates can be obtained by penalizing the likelihood by a functional which depends on the roughness of the logarithm of the density. The limiting case of the estimates as the amount of smoothing increasing has a natural form which makes the method attractive for data analysis and which provides a rationale for a particular choice of roughness penalty. The estimates are shown to be the solution of an unconstrained convex optimization problem, and mild natural conditions are given for them to exist. Rates of consistency in various norms and conditions for asymptotic normality and approximation by a Gaussian process are given, thus breaking new ground in the theory of maximum penalized likelihood density estimation.

AMS (MOS) Subject Classifications: Primary: 62G05
Secondary: 62E20, 46E35, 65D10, 65K10, 62-07

Key Words: probability density estimate, roughness penalty, penalized likelihood, smoothing, data analysis, reproducing kernel Hilbert space, Sobolev space, convex optimization, existence and uniqueness, rates, consistency, asymptotic normality, Gaussian process, strong approximation.

Work Unit Number 4 (Statistics and Probability)

*School of Mathematics, University of Bath, Claverton Down, Bath BA2 7AY, England

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

The basic problem considered is the estimation of a probability density function f given observations from the density. This problem arises virtually wherever data are collected, and is of particular interest in medical and engineering applications. Density estimates are useful for exploring properties of the data and for presenting data in a way comprehensible to the layman. They are also used for constructing versions of various statistical techniques (for example in automatic diagnosis and pattern recognition) which do not depend on specific assumptions about the underlying model.

A particular case of the estimates of the present paper is the following. Given data $X_1,\ldots,X_n$, choose the estimate $\hat{f}$ to maximize $n^{-1}\sum_{i=1}^{n} \log \hat{f}(X_i) - \frac{1}{2}\lambda \int \{(d/dx)^3 \log \hat{f}(x)\}^2 dx$, subject to $\int \hat{f} = 1$. The first term is, in a sense, the goodness of fit of $\hat{f}$ to the data, while the integral is a penalty for how 'wiggly' the estimate is. The parameter λ controls the amount by which the data are smoothed to obtain the estimate. As the parameter λ tends to infinity, the estimate converges to a normal density fitted to the data, and thus the definition of an 'infinitely smoothed' estimate is very natural; that is the advantage of the formulation of this paper over previous methods.

Because of their implicit definition, density estimates obtained in this way are rather intractable and little is known about their behavior. In this paper several new results on the large sample behavior of the estimates are obtained and so the work should help considerable in the understanding of roughness penalty procedures. In addition a characterization of the estimates as the solution to an unconstrained convex optimization procedure is given; apart from its mathematical value, this should be very useful when computing the estimates in practice.

ON THE ESTIMATION OF A PROBABILITY DENSITY FUNCTION BY THE MAXIMUM PENALIZED LIKELIHOOD METHOD

B. W. Silverman*

1. INTRODUCTION.

Good and Gaskins (1971) introduced the idea of roughness penalty density estimation. Their idea was to use as an estimate that density which maximized a penalized version of the likelihood. Given observations $X_1,\ldots,X_n$, the penalized log likelihood is defined as

$$\omega(f) = \sum \log f(X_i) - \alpha R(f)$$

where $R(f)$ is a 'flamboyancy functional' such as $\int (f'')^2$ and the parameter α controls the amount by which the data are smoothed to give the estimate. (Use the convention throughout that unqualified sums are over the range $i = 1$ to n.) Without the roughness penalty term the likelihood is unbounded above; intuitively the maximum likelihood estimator is a sum of delta function spikes at the observations. The Good-Gaskins formulation can be given a Bayesian justification; see their paper for details. An excellent exposition of penalized likelihood estimates is given by Tapia and Thompson (1978).

*School of Mathematics, University of Bath, Claverton Down, Bath BA2 7AY, England

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

In this paper a variation of the Good-Gaskins estimator is discussed. For compelling reasons given in Section 2 below, the logarithm of the density - rather than the density itself - will be penalized for roughness. In Section 3 it will be seen that the resulting constrained minimization can be replaced by an unconstrained convex optimization. Section 4 is concerned with conditions for existence of the estimator; these turn out to be mild and elegant.

In the remaining sections, the asymptotic properties of the estimator are discussed. Very little is known about the asymptotics of any roughness penalty methods of density estimation beyond the consistency (in a very strong norm, under quite restrictive conditions) proved by de Montricher (1981), and also the results for a related estimator proved by Reiss (1981). Some rates of convergence have been obtained by Klonias (1981); though his results are for different estimators than ours, they appear to be weaker insofar as a comparison is possible. For the estimators of this paper far more can be deduced. Sections 5 to 8 below lead to proofs of consistency with rates in a variety of different norms. In Section 6 a linear approximation is developed which is of considerable conceptual, as well as mathematical, value. The main consistency results are given in Section 8. The question of asymptotic normality is discussed in Section 9, where a uniform approximation of the estimator by a Gaussian process is given.

2. DEFINITION AND MOTIVATION.

Practically all density estimation methods have the property that the limiting estimate as the amount of smoothing decreases is a sum of spikes at the observations, but what happens as the amount of smoothing increases depends on exactly what method is being used. It turns out that roughness penalty estimates with a suitable penalty functional have a very attractive property, best illustrated by considering a special case. Suppose that the penalty

$$R_N(f) = \int_{-\infty}^{\infty} \{(d/dx)^3 \log f(x)\}^2 dx$$

is used. Then, in the sense made clear in Theorem 2.1 below, the limiting estimate as the parameter α tends to infinity will be the normal density with the same mean and variance as the data. Thus as α varies the method will give a range of estimates from the 'infinitely rough' sum of delta functions to the 'infinitely smooth' maximum likelihood normal fit to the data.

Computational and mathematical difficulties aside, this observation presents a very strong case for the use for density estimation of the roughness penalty method with penalty R_N. Since one of the objects of non-parametric methods is to investigate the effect of relaxing parametric assumptions, it seems sensible that the limiting case of a non-parametric density estimate should be a natural parametric estimate. These remarks also give a satisfying rationale for the choice of roughness functional. Previously this

choice has been made either in an ad hoc way or for reasons of mathematical or computational convenience.

Another advantage of this formulation is that the functional ω depends only on the logarithm of the density and so any density estimates obtained will automatically be positive. This remark is further elucidated below in Section 3 which deals with conditions under which the functional ω has a maximum. It should also be noted that the log density is itself a very natural quantity to escimate, particularly if the estimates are used to estimate likelihood functions, or for non-parametric discriminant analysis. Leonard (1978) has used a Bayesian approach to density estimation in which a stochastic process structure is placed on the log density; this differs from our approach both in its motivation and in some of its detail, but is nevertheless another example of penalizing for roughness in the logarithm of the density.

It is possible to define other roughness penalties according to other perceptions of 'infinitely smooth' exponential families of densities. The essential property, easily checked for the case discussed above, is that $R(f)$ should be zero if and only if f is in the required family. For example, for data on the half line, $R(f) = \int_0^\infty \{(\log f)''\}^2$ will give rise to exponential densities being the limiting case, while on the circle $R(f) = \int \{(\log f)''' + (\log f)'\}^2$ will have as the infinitely smooth family the von Mises densities defined by

$$f_{\kappa,\theta_0}(\theta) \quad \propto \exp\{\kappa \cos(\theta - \theta_0)\} ,$$

and discussed in detail by Mardia (1972).

We conclude this section with some definitions and the theorem which gives the form of the limiting estimates. Suppose that the domain of definition of the estimates and the set in which the observations lie is a connected set Ω in R^d. A space such as the circle is considered to be an interval in R^1 with periodic boundary conditions placed on all the functions considered; the imposition of these boundary conditions will not affect any of the results of this paper.

Suppose that D is a linear differential operator of the form

$$D(g) = \sum c\ (\alpha_1,\ldots,\alpha_d)\left(\frac{\partial}{\partial x_1}\right)^{\alpha_1}\ldots\left(\frac{\partial}{\partial x_d}\right)^{\alpha_d}$$

where the sum is over all vectors $\underset{\sim}{\alpha}$ of non-negative integers satisfying

$$1 \leq \sum \alpha_i \leq m$$

for some fixed integer m. Assume that at least one of the coefficients $c(\underset{\sim}{\alpha})$ for $\sum \alpha_i = m$ is non-zero. The results of Sections 2 and 3 will also hold where the coefficients $c(\underset{\sim}{\alpha})$ depend on x, but for the subsequent work it is assumed that there is no dependence of this kind. Note that there is no constant term in the definition of D; $D(g)$ depends only on derivatives of g. Define the non-negative definite bilinear form $[\ ,\ ]$ by

$$[g_1,g_2] = \int D(g_1)D(g_2) ;$$

here, and subsequently, unqualified integrals are taken to be over Ω with respect to Lebesgue measure.

Let S be the set of real functions g on Ω for which

(i) The $(m-1)^{th}$ derivative(s) of g exist everywhere and are piecewise differentiable.

(ii) $[g,g] < \infty$.

(iii) $\int e^g < \infty$.

Then given independent identically distributed observations $X_1,\ldots,X_n$ on Ω, our estimate $\hat{g}$ of the log density underlying the observations will be the solution, if it exists, of the constrained optimization problem

$$\max\{\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \frac{1}{2}\lambda[g,g]\}$$

subject to g in S and $\int e^g = 1$. The substitution $\lambda = 2\alpha/n$ has been made to simplify some of the mathematical expressions below. Our estimate $\hat{f}$ of the density itself is given by $\hat{f} = \exp(\hat{g})$.

The <u>null family</u> of the quadratic form $[\ ,\ ]$ will be defined to the collection of densities f on Ω for which $[\log f, \log f]$ is zero. It is easily shown that the null family is an exponential family, with at most $(m-1)$ parameters.

The following theorem gives a sense in which the 'infinitely smooth' estimator has the required form.

Theorem 2.1. Provided f is a density with $\log f$ in S, define

$$\omega_\lambda(f) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i) - \frac{1}{2}\lambda[\log f, \log f] .$$

Let f_∞ be the maximum likelihood estimator within the null family based on $X_1,\ldots,X_n$; suppose the data are such that f_∞ exists. Then, given any density $f \neq f_\infty$ with $\log f$ in S, for all sufficiently large λ

$$\omega_\lambda(f_\infty) > \omega_\lambda(f) .$$

Proof.

If f is not in the null family then $\omega_\lambda(f) \to -\infty$ as $\lambda \to \infty$ while $\omega_\lambda(f_\infty)$ remains fixed. If f is in the null family then

$$\omega_\lambda(f_\infty) - \omega_\lambda(f) = \frac{1}{n}\sum_i \{\log f_\infty(X_i) - \log f(X_i)\} > 0$$

by the definition of f_∞ as a maximum likelihood estimate. In either case the conclusion of the theorem holds, completing the proof.

Let f_λ denote the density (if it exists) which maximizes ω_λ; then it would be of interest to investigate further in what senses $f_\lambda \to f_\infty$ as $\lambda \to \infty$. We shall not consider this question further in this paper.

3. THE ESTIMATE AS AN UNCONSTRAINED OPTIMUM.

One of the reasons that roughness penalty density estimates present computational and mathematical difficulties is their

implicit definition as a solution to a constrained optimization problem. The results of this section show that our estimates can be found as the unconstrained minimum of a strictly convex functional without any unknown Lagrange multipliers. This observation has both mathematical and computational value, and is the foundation of the theoretical work given below. Computational aspects will not be considered here, but the result makes it possible to use standard methods for unconstrained convex problems to find the estimator; these will be explored in subsequent work.

For g in S and fixed $\lambda > 0$, write

$$A_0(g) = \frac{1}{2}\lambda[g,g] - \frac{1}{n}\sum g(X_i) \tag{3.1}$$

and

$$A(g) = \frac{1}{2}\lambda[g,g] + \int e^g - \frac{1}{n}\sum g(X_i) \ . \tag{3.2}$$

We show that the unconstrained minimum of $A(g)$ is identical with the constrained minimum of $A_0(g)$.

Theorem 3.1. The function $\hat{g}$ in S minimizes $A_0(g)$ over g in S subject to $\int e^g = 1$ if and only if $\hat{g}$ minimizes $A(g)$ over S.

Remarks. Note first that the theorem says nothing about the existence of $\hat{g}$; this question is considered in Section 4 below. Our reason for proving this result first is that we shall only deal with the existence question under conditions on Ω which are not needed for the present argument.

It is easily shown that A is a strictly convex functional on S, as defined on p. 154 of Tapia and Thompson (1978) and

hence, by their Theorem 2 on p. 160, it is an immediate corollary of the present theorem that $\hat{g}$ is unique if it exists at all.

Proof of the Theorem.

Given g in S, define $g^* = g - \log \int e^g$, so that $\int \exp(g^*) = 1$. Since $[\ ,\ ]$ only involves derivatives, $[g^*,g^*] = [g,g]$. Therefore it follows by elementary manipulations that $A(g^*) = A(g) + 1 - \int e^g + \log \int e^g$ and so $A(g^*) \leq A(g)$ with equality only if $\int e^g = 1$, since $t - \log t > 1$ for all $t > 0$, with equality only if $t = 1$. Therefore, $\hat{g}$ minimizes $A(g)$ if and only if $\hat{g}$ minimizes $A(g)$ subject to $e^g = 1$; but subject to $\int e^g = 1$, $A(g)$ and $A_0(g) + 1$ are identical, and so the proof of the theorem is complete.

Note that the proof of the theorem depends crucially on the fact that the penalizing functional involves only derivatives.

4. EXISTENCE OF THE ESTIMATORS.

A discussion of the existence properties of the Good-Gaskins estimators is given in Chapter 4 of Tapia and Thompson (1978), drawing on material from de Montricher, Tapia and Thompson (1975). It is clear from that work both that the estimators defined in this paper cannot be shown to exist by existing work, and also that the question of existence can be a little delicate.

The theorem of this section gives a natural and elegant condition for the existence of the estimates. For convenience the

theorem is stated for the special case of univariate bounded Ω, but remarks about generalizations are made below.

Theorem 4.1. Suppose Ω is a bounded interval in R^1, possibly subject to periodic boundary conditions. Given observations $X_1,\ldots,X_n$ in Ω, the functional A as defined in (3.2) above will have a minimizer in S if there is a maximum likelihood estimator based on $X_1,\ldots,X_n$ in the null family.

Remarks. The condition of the existence of a maximum likelihood estimate in the null class is, of course, a very mild one. In the case where Ω is the circle and the null class is the von Mises family, for example, all that is required is at least two distinct data points. It is interesting to compare the existence condition to the condition given in a different context for the existence of the estimator considered by Silverman (1978b); it is presumably possible to extend the technique of this proof to deal with penalized likelihood estimators of quantitites other than probability density functions.

Proof.

The proof depends on properties of reproducing kernel Hilbert spaces; see, for example, Oden and Reddy (1976) for an account of these. Given g_1 and g_2 in S, define

$$\langle g_1, g_2 \rangle = [g_1, g_2] + \int g_1 g_2 \qquad (4.1)$$

and

$$\|g_1\|_0 = \langle g_1, g_1 \rangle_0^{1/2}.$$

Since Ω is bounded, the norm $\|\cdot\|_0$ will make S a reproducing kernel Hilbert space equivalent to the Sobolev space $H^m(\Omega)$.

Define subspaces S_1 and S_2 of S by

$$S_1 = \{g \text{ in } S: [g,g] = 0 \text{ and } \int g = 0\}$$

$$S_2 = \{g \text{ in } S: \int g = 0 \text{ and } \langle g, g_1 \rangle = 0 \text{ for all } g_1 \text{ in } S_1\}$$

If ρ is the largest eigenvalue less than one of the reproducing kernel in S, then, given g_2 in S_2,

$$\|g_2\|_0^2 \geq \rho^{-1} \int g_2^2$$

and hence

$$[g_2, g_2] \geq (1 - \rho^{-1}) \|g_2\|_0^2$$

Since Ω is bounded, by the conditions imposed on m and Ω, the Sobolev embedding theorem implies that the sup norm is continuous with respect to $\|\cdot\|_0$ and hence there is a constant C such that, for g_2 in S_2,

$$\sup|g_2| \leq C[g_2, g_2]^{1/2}. \tag{4.2}$$

Define spaces S_0, S^* and S_1^* by

$$S_0 = \{g \text{ in } S: \int g = 0\},$$

$$S^* = \{g \text{ in } S: \int e^g = 1\},$$

$$S_1^* = \{g \text{ in } S: \int e^g = 1 \text{ and } [g,g] = 0\}.$$

Define a functional A^* by, given g in S, defining A_0 as in (3.1),

$$A^*(g) = A_0(g) + \log \int e^g.$$

It is easily shown that $A^*(g) = A^*(g + c)$ for all constants c and hence that the mappings $g \to g - \log \int e^g$ and $g \to g - \int g$ set up an A^* preserving (1-1) correspondence between S_0 and

S^*. Since A_0 and A^* coincide on S^*, it follows that A_0 will have a minimum in S^* if and only if A^* has a minimum in S_0. A minimum of A_0 on S^* is, of course, precisely the estimate we are seeking; therefore it will suffice to show that there is a minimum of A^* on S_0.

Given any g in S_0, write $g = g_1 + g_2$ with g_1 in S_1 and g_2 in S_2. Then

$$A^*(g) = \frac{1}{2}\lambda[g,g] + \log \int e^g + 1 - \frac{1}{n}\sum g(X_i)$$

$$> \frac{1}{2}\lambda[g_2,g_2] - \frac{1}{n}\sum g_2(X_i) \qquad (4.3)$$

$$+ \log\{\exp(\inf g_2) \int e^{g_1}\} + 1 - \frac{1}{n}\sum g_1(x_1)$$

using the fact that $[g,g] = [g_2,g_2]$. From (4.3) it follows that

$$A^*(g) > \frac{1}{2}\lambda[g_2,g_2] - \frac{1}{n}\sum g_2(X_i) \qquad (4.4)$$

$$+ \inf g_2 + A^*(g_1) + 1 .$$

The (1-1) correspondence defined above between S_0 and S^* gives an A^* preserving correspondence between S_1 and S_1^*. On S_1^*, A^* is precisely $-\frac{1}{n}\sum g(X_1)$, and so the log density $\hat{g}$ of the maximum likelihood estimator within the null class will be a minimizer of A^* in S_1^*; it follows that $\hat{g} - \int \hat{g}$ will be a minimizer of A^* in S_1. By Cauchy-Schwarz it is easily shown that A^* is strictly convex on S_1, and hence, since S_1 is a finite dimensional normed space on which A^* has a minimum, it can be shown by elementary functional analysis that there exist

constants $C_1 > 0$ and C_2 such that, for g_1 in S_1,

$$1 + A^*(g_1) \geq C_1 \|g_1\|_0 + C_2 \,. \tag{4.5}$$

Next we consider the terms of (4.4) involving g_2. Using inequality (4.2) it follows that, for fixed $\lambda < 0$, there exist positive constants C_3 and C_4 such that

$$\begin{aligned} &\tfrac{1}{2}\lambda\,[g,g] - \frac{1}{n}\sum g_2(x_1) + \inf g_2 \\ &\qquad \geq \tfrac{1}{2}\lambda[g_2,g_2] - 2\sup|g_2| \\ &\qquad \geq C_3\|g_2\|_0^2 - C_4\|g_2\|_0 \end{aligned} \tag{4.6}$$

Substituting (4.5) and (4.6) into (4.4) gives, for g in S_0,

$$\begin{aligned} A^*(g) &\geq C_1\|g_1\|_0 + C_2 + C_3\|g_2\|_0^2 - C_4\|g_2\|_0 \\ &\geq C_5(\|g_1\|_0 + \|g_2\|_0) + C_6 \end{aligned} \tag{4.7}$$

for suitable constants $C_5 > 0$ and C_6, by elementary algebra. From (4.7), using the triangle inequality,

$$A^*(g) \geq C_5\|g\|_0 + C_6 \,. \tag{4.8}$$

By Theorem 5, p. 162 of Tapia and Thompson (1978), it follows that A^* has a minimizer on S_0, completing the proof of the theorem.

The extension of the theorem to the case where Ω is a bounded multivariate domain is straightforward provided that the supremum operator is continuous with respect to the norm $\|\cdot\|_0$ on S; this will entail conditions on Ω and on D, for details of which see a text on Sobolev spaces. An extension to unbounded

Ω will require a different technique of proof since it will no longer necessarily be the case that $\int g^2 < \infty$ for g in S.

5. ASYMPTOTIC PROPERTIES - PRELIMINARIES.

In the remaining sections, the consistency and other asymptotic properties of the estimators are studied. There has been very little work on the consistency properties of maximum penalized likelihood density estimates; the main contribution is the paper of de Montricher (1981), whose results do not seem to be directly applicable to our estimates and who does not consider questions of rates of consistency or of asymptotic distributions. See also Klonias (1981). Consistency of a related class of estimators has also been considered by Reiss (1981). The techniques used in this paper are more akin to those used in several papers of Wahba (e.g. Wahba, 1977) though some care is needed because the functional A, though unconstrained, is not quadratic.

For the remainder of the paper attention will be restricted to the case where Ω is a bounded univariate domain, possibly with periodic end conditions. The extension to any particular multivariate case will depend on the solution to the eigenvalue problem of the differential operator D in the domain Ω; once that is solved the arguments of this paper will go through easily.

It will be assumed that the observations $X_1,\ldots,X_n$ are independent and identically distributed with density f_0 on Ω.

In order to make what is quite an involved argument a little more transparent, rather stringent smoothness conditions will be placed on f_0, but it should be stressed that appropriate versions of the theorems remain true under much milder assumptions, and can be obtained by very similar techniques. These extensions are left to the reader to investigate.

Let $g_0 = \log f_0$. Assume throughout that, in the terminology of Wahba (1977), g_0 is very smooth, in other words that g_0 and its periodic extension have 2m derivatives on Ω and $\int \left(g_0^{(2m)}\right)^2$ is finite. In particular, assume that g_0 is bounded above and below. It will be convenient to prove results about the convergence of the estimates of the log density rather than the density itself, but only elementary calculus is needed to transform these back to results about the density.

The minimizer of the functional A of (3.2) will be denoted by $\hat{g}$. The explicit dependence of $\hat{g}$ and related quantities on the sample size, the values of the observations, and the smoothing parameter will usually be suppressed, as will the dependence of the smoothing parameter on the sample size. The basic strategem of the consistency proof is first to study the properties of a function g_1 (defined in Section 6 below) which is a linear approximation to $\hat{g}$, and then to show that $\hat{g}$ and g_1 are sufficiently close to allow results for $\hat{g}$ to be obtained. It should be kept in mind throughout that although g_1 has desirable properties which help one understand the behavior of $\hat{g}$, the definition of g_1 depends

on the unknown density f_0; therefore g_1 is only a mathematical device and cannot, in contrast to $\hat{g}$, be calculated in practice.

The remainder of this section consists of definitions and lemmas which set up the technical machinery needed in the subsequent sections. A first reader may find it easier to skip to Section 6 and then to refer back as necessary. A more casual reader could skip straight to Section 8, where the main results are given.

Three different norms will be used in the study of consistency; these will be defined for g in S as follows:

$$\|g\|_2^2 = \int g^2 f_0 \ ;$$

$$\|g\|_\infty^2 = \sup_\Omega |g| \ ; \tag{5.1}$$

$$\|g\|_S^2 = [g,g] + \int g^2 f_0 \ .$$

Inner products $\langle \ \rangle_2$ and $\langle \ \rangle_S$ are defined by

$$\langle g_1, g_2 \rangle_2 = \int g_1 g_2 f_0 \ ;$$

$$\langle g_1, g_2 \rangle_S = [g_1, g_2] + \int g_1 g_2 f_0 \ .$$

Since f_0 is bounded above and below away from zero, the norm $\| \ \|_S$ is equivalent to the Sobolev norm on $S = H^m(\Omega)$. Suppose $\{\phi_\nu : \nu > 0\}$ is a sequence of orthonormal eigenfunctions with respect to the density f_0 of the reproducing kernal of $\langle \ , \ \rangle_S$; i.e., for a sequence of eigenvalues $\{\lambda_\nu\}$

$$\langle \phi_i, \phi_j \rangle_S = \lambda_i^{-1} \delta_{ij}$$

and

$$\langle \phi_i, \phi_j \rangle_2 = \delta_{ij}$$

where δ_{ij} is the Kronecker delta. By standard arguments (see, for example, Riesz and Nagy, 1955), ϕ_0 is identically equal to 1 and the eigenvalues satisfy

$$1 = \lambda_0 > \lambda_1 > \lambda_2 > \cdots .$$

Define the sequence ρ_ν by

$$\rho_\nu = \lambda_\nu^{-1} - 1 ;$$

then it is immediate that

$$[\phi_i, \phi_j] = \rho_i \delta_{ij} .$$

When expanding elements of S in terms of the eigenfunctions, we shall use an additional subscript (enclosed in brackets if any confusion between subscripted functions and coefficients may arise) to denote a coefficient. Thus, for example, we shall write

$$g_0 = \sum g_{0\nu}\phi_\nu = g_{00}\phi_0 + g_{01}\phi_1 + \cdots$$

$$g = \sum g_\nu \phi_\nu = g_{(0)}\phi_0 + g_{(1)}\phi_1 + \cdots .$$

Unqualified sums over ν will be taken to be over the range $\nu = 0$ to ∞. The asymptotic behavior of the eigenvalues can be deduced using the following lemma, which says that replacing f_0 by the constant function does not affect the rate of convergence of the eigenvalues to zero.

Lemma 5.1. Suppose that the sequences λ_ν and ϕ_ν are defined as above, and suppose that λ_ν^* are the eigenvalues of the L^2

orthonormal eigenfunctions ϕ_ν^* of the inner product $[\ ,\ ]_0$ as defined in (4.1) above. Then, for all $\nu > 0$, putting

$$\rho_\nu = \lambda_\nu^{-1} - 1 \quad \text{and} \quad \rho_\nu^* = \lambda_\nu^{*-1} - 1,$$

$$\rho_\nu \inf f_0 \leq \rho_\nu^* \leq \rho_\nu \sup f_0 . \tag{5.2}$$

Proof.

The eigenvalue ρ_ν will satisfy

$$\rho_\nu = \inf\{[g,g] : \int g\phi_j f_0 = 0,\ j = 0,\ldots,\nu-1, \text{ and } \int g^2 f_0 \geq 1\}$$

$$\leq \inf\{[g,g] : \int g\phi_j f_0 = 0 \text{ and } \int g^2 = (\inf f_0)^{-1}\} \tag{5.3}$$

since the infimum is over a smaller set. By an analogous argument to Riesz and Nagy (1955) p. 238, it follows from (5.3) that, taking the supremum over elements $h_0,\ldots,h_{\nu-1}$ of S,

$$\rho_\nu \leq \sup_{h_0,\ldots,h_{\nu-1}} \inf_g \{[g,g] : \int g h_j = 0,\ j = 0,\ldots,\nu-1$$

$$\text{and } \int g^2 = (\inf f_0)^{-1}\}$$

$$= (\inf f_0)^{-1} \rho_\nu^* .$$

The other inequality of (5.2) is proved similarly, completing the proof.

Corollary. There exist positive constants α and β such that, for all $\nu > 0$,

$$\lambda_\nu = c_\nu \nu^{-2m}, \quad \alpha \leq c_\nu \leq \beta .$$

Proof.

By standard properties of Sobolev spaces, it is easily shown that the L^2 orthonormal eigenfunctions of the inner product $\langle\ ,\ \rangle_0$ have eigenvalues λ_ν^* which decay exactly at rate ν^{-2m}. To see this, note that the eigenfunction expansion is precisely a Fourier series expansion and that the eigenvalues are reciprocals of polynomials in ν of degree 2m. Now apply Lemma 5.1 to obtain the rate of decay of λ_ν.

Given any g in $L^2(\Omega)$, it is now possible to give expressions for the norms of (5.2) in terms of the coefficients of the eigenfunction expansion of g.

Lemma 5.2. Suppose g is in $L^2(\Omega)$ and $g_\nu = \int g\phi_\nu f_0$. Then

$$\|g\|_2^2 = \sum g_\nu^2 , \tag{5.4}$$

$$\|g\|_S^2 = \sum \lambda_\nu^{-1} g_\nu^2 , \tag{5.5}$$

and, given $\varepsilon > 0$, there exists $C_\varepsilon > 0$ such that

$$\|g\|_\infty^2 \leq C_\varepsilon \sum \nu^{1+\varepsilon} g_\nu^2 . \tag{5.6}$$

Furthermore, if g is very smooth, then

$$\sum \lambda_\nu^{-2} g_\nu^2 < \infty . \tag{5.7}$$

Proof.

Equations (5.4) and (5.5) are immediate from the definition of the eigenfunctions and eigenvalues. The equations (5.6) and (5.7) follow by considering Hilbert scales of spaces as considered, for example, by Oden and Reddy (1976) Chapter 4. From there (p. 133) and the corollary above it follows that $\sum \nu^{1+\varepsilon} g_\nu^2$ will be equivalent to the norm of the fractional Sobolev space of index $\frac{1}{2}(1+\varepsilon)$, and hence, by the Sobolev embedding theorem (see the remarks on p. 109 of Oden and Reddy) the inequality (5.6) follows at once. Similarly the norm $\sum \lambda_\nu^{-2} g_\nu^2$ is equivalent to the $H^{2m}(\Omega)$ Sobolev norm, proving the last part of the lemma.

The next lemma gives the asymptotic behaviour of certain functions of the eigenvalues which will occur in Section 6 below. The notation $f_1(\lambda) \sim f_2(\lambda)$ as $\lambda \to 0$ is taken to mean that $f_1(\lambda)/f_2(\lambda)$ and $f_2(\lambda)/f_1(\lambda)$ are bounded as $\lambda \to 0$. The lemma is a generalization of the estimates obtained by Wahba (1977) for her $A(\lambda)$ and $G(\lambda)$.

Lemma 5.3. Given $a < 4m - 1$, as $\lambda \to 0$,

$$\sum_{\nu=1}^{\infty} \frac{\nu^a}{(1 + \lambda\rho_\nu)^2} \sim \lambda^{-(a+1)/2m}$$

and, provided g_0 is very smooth, as $\lambda \to 0$, given $b < 2m$,

$$\sum_{\nu=1}^{\infty} \frac{\nu^a}{(1 + \lambda\rho_\nu)^2} \begin{cases} \to \sum \rho_\nu^2 g_{0\nu}^2 & \text{if } b = 0 \\ = o(\lambda^{-b/2m}) & \text{if } b > 0 \end{cases}$$

Proof.

The first part is proved using the corollary to Lemma 5.1 by approximating the sum by an integral in the manner of Wahba (1977) p. 660; values for the implied constants can be obtained by more careful analysis. The second part is obtained by an application of the dominated convergence theorem. Note that making milder smoothness assumptions on g_0 will affect the rates in the second part of the lemma. Some care is necessary if Ω is not a periodic domain though; see Rice and Rosenblatt (1981). The final part of this section concerns the sample coefficients of the empirical distribution. Define a random sequence β_ν by

$$\beta_0 = 0$$

$$\beta_\nu = n^{-1} \sum_{i=1}^{n} \phi_\nu(X_i) \quad \text{for} \quad \nu > 0 .$$

The β_ν depend on n, but this dependence will not be expressed explicitly. Some properties of the β_ν are given in the following lemma, the proof of which follows immediately from the facts that $\phi_0 = 1$ and that the ϕ_ν are orthonormal with respect to f_0.

Lemma 5.4. Given n, the sequence β_ν satisfies

$$E\beta_r = 0 \quad \text{for all} \quad r; \quad \text{and}$$

$$E\beta_r\beta_s = n^{-1}\delta_{rs} \quad \text{for} \quad r \quad \text{and} \quad s > 1 .$$

It is now immediate, by the classical central limit theorem, that for each $\nu > 0$, $n^{1/2}\beta_\nu$ will have, asymptotically, a standard

normal distribution. Indeed it is possible to provide a simultaneous strong approximation of the sequence β_ν by a sequence of normal random variables; this is done in the following lemma, which is the last result of this section.

Lemma 5.5. On a suitable probability space, defining the sequence β_ν as above, there exists for each n a sequence $\tilde{\beta}_\nu$ of independent $N(0,1)$ random variables such that, with probability 1,

$$\limsup_{n\to\infty} n^{1/2}(\log n)^{-1} \sup_{\nu\geq 1} \nu^{-1} |n^{1/2}\beta_\nu - \tilde{\beta}_\nu| \leq C(f_0)$$

where $C(f_0)$ is a constant depending only on f_0.

Proof.

Write $\beta_\nu = \int \phi_\nu(t)dF_n(t)$ where F_n is the empirical distribution function of the observations and then proceed as in the proof of Propositions 1 and 2 of Silverman (1978a), approximating $n^{1/2}(F_n - F)(t)$ by a transformed Brownian bridge $W_n^0\{F(t)\}$ using Theorem 3 of Komlos, Major and Tusnady (1975). The $\tilde{\beta}_\nu$ are the coefficients of the expansion of $W_n^0\{F(t)\}$ in terms of the eigenfunctions and are easily shown to have the required structure. Defining $Z_n(t)$ as in Silverman (1978a) it follows as on p. 179 of that paper that, using the fact that $\int \phi_\nu f_0$ is zero for $\nu > 1$,

$$n(\log n)^{-1} |n^{-1/2}\beta_\nu - \tilde{\beta}_\nu| \leq \int |\phi_\nu'| \sup |Z_n(t)| . \qquad (5.8)$$

For fixed f_0, by p. 134 of Oden and Reddy (1976) there are constants C_1, $C_2(f_0)$ such that

$$\begin{aligned} \int |\phi_\nu'| &\leq C_1 \{\int (\phi_\nu')^2\}^{1/2} \\ &\leq C_2(f_0) \|\phi_\nu\|_2^{(m-1)/2m} \|\phi_\nu\|_s^{1/2m} \\ &= C_2(f_0) \lambda_\nu^{-1/2m} = 0(\nu) \end{aligned} \tag{5.9}$$

by the corollary to Lemma 5.1. To complete the proof substitute (5.9) and (2) of Silverman (1978a) into (5.8).

6. THE LINEAR APPROXIMATION.

In this section the linear approximation g_1 to g will be defined and studied; the question of the closeness of the approximation will be considered in Section 7 below. The approximation is linear in the sense that it is a linear function *of the transformed observations* $\phi_\nu(X_i)$ and that it is the solution of a certain linear system in a Hilbert space. It is this linearity which leads to the tractability of the approximation.

Define a quadratic form A_1 for g in S by

$$A_1(g) = \frac{1}{2} \lambda[g,g] + \int \{1 + (g-g_0) + \frac{1}{2} (g-g_0)^2\} f_0 - \frac{1}{n} \sum_{i=1}^{n} g(X_i) \tag{6.1}$$

The motivation behind the definition of A_1 is that it is the quadratic form which has second order contact with the functional

A at g_0. Furthermore, by Proposition 17 and Theorem 6 of Appendix I of Tapia and Thompson (1978), the functional A_1 is uniformly convex on S and hence has a unique minimizer g_1 in S.

Though the function g_1 is, like $\hat{g}$, defined implicitly, it is straightforward to write its eigenfunction expansion explicitly. Up to a constant, we have, for g in S,

$$\begin{aligned} A_1(g) &= \frac{1}{2}\lambda \sum \rho_\nu g_\nu^2 + g(0) + \frac{1}{2}\sum g_\nu^2 \\ &\quad - \sum g_\nu g_{0\nu} - \sum_{\nu=0}^{\infty}\sum_{i=1}^{n} n^{-1} g_\nu \phi_\nu(X_i) \qquad (6.2) \\ &= \frac{1}{2}\sum (\lambda\rho_\nu + 1) g_\nu^2 - \sum (g_{0\nu} + \beta_\nu) g_\nu \end{aligned}$$

where we have used the fact that $n^{-1}\sum \phi_0(X_i) = 1$. It follows from (6.2) that the coefficients of g_1 satisfy

$$g_{1\nu} = \frac{g_{0\nu} + \beta_\nu}{1 + \lambda\rho_\nu} \qquad (6.3)$$

Studying these coefficients gives several asymptotic results for $g_0 - g_1$. Notice that the form (6.3) can immediately be decomposed into its systematic and random components, so that

$$E(g_{1\nu} - g_{0\nu}) = \lambda\rho_\nu g_{0\nu}(1 + \lambda\rho_\nu)^{-1} \qquad (6.4)$$

and

$$g_{1\nu} - Eg_{1\nu} = \beta_\nu(1 + \lambda\rho_\nu)^{-1} \tag{6.5}$$

Consideration of (6.4) and (6.5) as $\lambda \to 0$ shows that there is, as in most smoothing problems, a trade off between bias and random error.

It is very straightforward to apply the results of Section 5 to give asymptotic properties of g_1, and this is done in the following theorem. Both the uniform and the L^2 rates of convergence will be required in Section 7, while the Sobolev rate is included for its own interest.

Theorem 6.1. Defining g_1 as above, and using the definitions of Section 5 for the various norms, as $\lambda \to 0$ and $n \to \infty$,

$$E\|g_1 - g_0\|_2^2 \sim n^{-1}\lambda^{-1/2m} + \lambda^2$$

$$E\|g_1 - g_0\|_S^2 = O(n^{-1}\lambda^{-(2m+1)/2m}) + O(\lambda)$$

and, given $\delta > 0$,

$$E\|g_1 - g_0\|_\infty^2 = o\{\lambda^{-\delta}(n^{-1}\lambda^{-1/m} + \lambda^{(4m-1)/2m})\}$$

Proof.

From Lemma 5.2 and (6.3) it follows that

$$\|g_1 - g_0\|_2^2 = \sum (g_{1\nu} - g_{0\nu})^2 = \sum \frac{(-\lambda\rho_\nu g_{0\nu} + \beta_\nu)^2}{(1 + \lambda\rho_\nu)^2}$$

and hence, by Lemma 5.4

$$E\|g_1 - g_0\|_2^2 = \sum_{\nu=0}^{\infty} \frac{\lambda^2\rho_\nu^2 g_{0\nu}^2}{(1 + \lambda\rho_\nu)^2} + \frac{1}{n}\sum_{\nu=1}^{\infty} \frac{1}{(1 + \lambda\rho_\nu)^2}$$

Substituting the bounds given by Lemma 5.3 completes the proof of the first part of Theorem 6.1. The second and third parts are proved in exactly the same way, by first applying Lemma 5.2 to give a bound for the appropriate norm of $(g_1 - g_0)$ in terms of the coefficients $g_{1\nu} - g_{0\nu}$, and then applying Lemmas 5.4 and 5.3.

It is easy to deduce conditions under which g_1 will converge to g_0 in various norms. In addition optimal rates of convergence can be obtained; these will be discussed further after it has been shown that, under suitable conditions, $\|g - g_1\|$ can be neglected relative to $\|g_1 - g_0\|$.

7. CLOSENESS OF THE LINEAR APPROXIMATION TO THE TRUE MINIMIZER.

In this section the closeness of g_1 to $\hat{g}$ will be considered. The arguments are a little involved, mainly because the functional A, while being strictly convex, is not uniformly convex. The major part of the section is taken up with the proof of the following lemma; at the end of the section a corresponding result for the Sobolev norm is discussed. The notation 0_p denotes an order of magnitude in probability.

Lemma 7.1. *Suppose the definitions and conventions of Sections 5 and 6 are used, and that* $\lambda \to 0$ *and* $n^{m-\delta}\lambda \to \infty$ *for some* $\delta > 0$ *as* $n \to \infty$. *Then, for all sufficiently small* $\varepsilon > 0$, *as* $n \to \infty$,

$$\|\hat{g} - g_1\|_\infty = 0_p(\lambda^{-\varepsilon}\{n^{-1}\lambda^{-1/m} + \lambda^{(4m-1)/2m}\}) .$$

Proof.

The proof of the lemma proceeds in several stages. First a new approximation $\hat{g}_M$ to $\hat{g}$ is defined, for which it is the case that the uniform convergence of $\hat{g}_M$ to g_0 will imply that $\hat{g}$ and $\hat{g}_M$ are eventually identical. The function $\hat{g}_M$ is the minimizer of a functional A_M; the derivative of A_M at g_1 can be bounded in such a way as to enable rates of stochastic convergence to zero of $\sup|\hat{g}_M - g_1|$ to be obtained, and these rates are easily shown to apply to $\sup|\hat{g} - g_1|$ also. Choose a number M such that

$$\sup|g_0| + 2 < M$$

and define the function $\exp_M$ by

$$\exp_M(x) = \begin{cases} \{1 + (x + M) + \frac{1}{2}(x + M)^2\}e^{-M} & \text{for } x < -M ; \\ e^x & \text{otherwise.} \end{cases}$$

Define a functional A_M on S by

$$A_M(g) = \frac{1}{2}\lambda[g,g] + \int \exp_M(g) - \frac{1}{n}\sum_{i=1}^{n} g(X_i) ,$$

and let $\hat{g}_M$ denote the minimizer of A_M; this functional is easily shown to be uniformly convex as defined by Tapia and Thompson (1978), and hence $\hat{g}_M$ exists and is unique.

In the remainder of the argument, derivatives of functionals will be used; these are Gateaux derivatives as discussed by Tapia and Thompson (1978). Note first that A_M and A and hence their first and second derivatives agree if $\sup|g| < M$; it is easy to

show that these derivatives exist everywhere. By the strict convexity of A and A_M, their respective minima correspond exactly to zeros of A' and A'_M and hence $\hat{g}$ and $\hat{g}_M$ will be equal if $\sup|\hat{g}_M| < M$.

Defining A_1 as in (6.1), since g_1 minimizes A_1 it will be the case that $A'_1(g_1)$ is zero, in other words, for all u in S,

$$\lambda[g_1,u] + \int u\{1 + (g_1-g_0)\}f_0 - n^{-1} \sum u(X_i) = 0 . \quad (7.1)$$

Now, substituting (7.1), we will have

$$\begin{aligned} A'(g_1)(u) &= \lambda[g_1,u] + \int u \exp(g_1) - n^{-1} \sum u(X_i) \\ &= \int u[\exp(g_1) - \{1 + (g_1-g_0)\}\exp(g_0)] . \end{aligned} \quad (7.2)$$

By elementary analysis, there exists a constant C such that, <u>provided</u> $\sup|g_1 - g_0| < 1$,

$$\begin{aligned} |A'(g_1)(u)| &\leq C \int |u|(g_1 - g_0)^2\exp(g_0) \\ &\leq C \sup|u| \, \|g_1 - g_0\|_2^2 ; \end{aligned} \quad (7.3)$$

by standard functional analysis, using the operator norms corresponding to those defined in (5.1) above,

$$\|A'_M(g_1)\|_\infty \leq C \|g_1 - g_0\|_2^2 . \quad (7.4)$$

Since, under the assumption $\sup|g_1 - g_0| < 1$, it follows <u>a fortiori</u> that $\sup|g_1| < M$ and hence $A'_M(g_1) = A'(g_1)$.

Now consider the operator $A''_M(g)$. Given any u in S,

$$A''_M(g)(u,u) = \lambda[u,u] + \int u^2 \exp''_M(g) \geq \lambda[u,u] + e^{-2M} \int u^2 f_0$$

$$= \sum_\nu (\lambda\rho_\nu + e^{-2M}) u_\nu^2 = \sum_\nu \nu^{-1-\varepsilon}(\lambda\rho_\nu + e^{-2M})\nu^{1+\varepsilon} u_\nu^2$$

for any $\varepsilon > 0$. By elementary analysis it follows that

$$A''_M(g)(u,u) \geq C^0_{M,\varepsilon} \lambda^{(1+\varepsilon)/2M} \sum \nu^{1+\varepsilon} u_\nu^2$$

$$\geq C_{M,\varepsilon} \lambda^{(1+\varepsilon)/2m} (\sup|u|)^2 \tag{7.5}$$

for suitable positive constants $C^0_{M,\varepsilon}$ and $C_{M,\varepsilon}$, by Lemma 5.2.

Now set $u_M = g_1 - \hat{g}_M$. Apply Taylor's theorem to the function (of t) $A'_M(\hat{g}_M + tu_M)(u_M)$ to obtain, for some θ, $0 < \theta < 1$,

$$A'_M(g_1)(u_M) = A''_M(\hat{g}_M + \theta u_M)(u_M, u_M) \ . \tag{7.6}$$

Combining (7.4), (7.5) and (7.6) it follows that

$$C_{M,\varepsilon} \lambda^{(1+\varepsilon)/2m} (\sup|u_M|)^2 \leq C \sup|u_M| \, \|g_1 - g_0\|_2^2$$

so that, for a suitable constant C_1, provided $\sup|g_1 - g_0| < 1$,

$$\sup|u_M| \leq C_1 \lambda^{-(1+\varepsilon)/2m} \|g_1 - g_0\|_2^2 \ . \tag{7.7}$$

Under the conditions stated in Theorem 7.1, it follows from Theorem 6.1 that

$$\sup|g_1 - g_0| \to 0 \quad \text{in probability} \tag{7.8}$$

and hence

$$P(\sup|g_1 - g_0| < 1) \to 1 \quad \text{as} \quad n \to \infty. \tag{7.9}$$

Again from Theorem 6.1 we have

$$\|g_1 - g_0\|_2^2 = O_p(n^{-1}\lambda^{-1/2m} + \lambda^2) \tag{7.10}$$

Combining (7.7), (7.9) and (7.10) gives, for all $\varepsilon > 0$,

$$\sup|g_1 - \hat{g}_M| = O_p(\lambda^{-\varepsilon}\{n^{-1}\lambda^{-1/m} + \lambda^{(4m-1)/2m}\}) . \tag{7.11}$$

In particular it follows that $\sup|g_1 - \hat{g}_M| \to 0$ in probability; combined with (7.8) this implies that $\sup|\hat{g}_M - g_0| \to 0$ in probability, from which it is immediate that

$$P(\sup|\hat{g}_M| < M) \to 1 \quad \text{as} \quad n \to \infty ;$$

by the remarks made near the beginning of the proof this implies that

$$P(\hat{g}_M \neq \hat{g}) \to 0 \tag{7.12}$$

and hence the proof of the lemma follows from (7.11).

No attempt will be made to obtain a finer bound for $\|\hat{g} - g_1\|_2$ since the bound obtained from Lemma 7.1 will suffice. A bound for the difference between g and g_1 in the Sobolev norm is given in the following lemma.

Lemma 7.2. Under the same conditions as Lemma 7.1,

$$\|\hat{g} - g_1\|_S = O_p(n^{-1}\lambda^{-(2m+1)/2m} + \lambda)$$

as n tends to infinity.

Proof.

The argument is very similar to the proof of Lemma 7.1. Since the sup operator is continuous in the Sobolev norm it follows from (7.4) that, for a suitable constant C_S, provided $\sup|g_1 - g_0| < 1$,

$$\|A'_M(g_1)\|_S < C_S \|g_1 - g_0\|_2^2 .$$

By an argument similar to that used to demonstrate (7.5), using Lemma 5.2, for all g and u in S we have, for a suitable constant C_M^S,

$$A''_M(g)(u,u) > C_M^S \lambda \|u\|_S^2 .$$

It can now be deduced that, provided $\sup|g_1 - g_0| < 1$

$$\|u_M\|_S = O(\lambda^{-1}) \|g_1 - g_0\|_2^2 ;$$

the remainder of the proof, making use of (7.12), exactly parallels that of Lemma 7.2.

8. THE MAIN CONSISTENCY RESULTS.

It is now possible to state and prove conditions under which $\hat{g}$ is in various senses a consistent estimator of g_0 and to give rates for this consistency. These are given in the following theorem. It should be stressed again that the conditions, particularly those placed on the smoothness of g_0, can be

weakened considerably by extending the arguments used in Sections 5, 6 and 7 above.

Theorem 8.1. Suppose Ω is a bounded univariate domain, possibly with periodic end conditions. Suppose the true density f_0 on Ω is bounded above and below away from zero; let $g_0 = \log f_0$. Suppose the roughness penalty $[g,g]$ is defined, using a differential operator of order m, as in Section 2 above, and that the log density estimate $\hat{g}$ is defined as in Section 3 above, based on independent identically distributed observations $X_1,\ldots,X_n$ from f_0. Suppose that $\int_\Omega (g_0^{(2m)})^2 < \infty$ and that $g_0^{(2m-1)}$ is continuous on the periodic extension of Ω.

Suppose throughout that the smoothing parameter λ satisfies, for some $\delta > 0$,

$$\lambda \to 0 \text{ and } n^{m-\delta}\lambda \to \infty \text{ as } n \to \infty .$$

Then $\hat{g}$ is uniformly consistent as an estimator of g_0 and in addition, for all $\varepsilon > 0$,

$$\sup_\Omega |\hat{g} - g_0|^2 = o_p\{\lambda^{-\varepsilon}(n^{-1}\lambda^{-1/m} + \lambda^{(4m-1)/2m})\} .$$

If, in addition, $n^{(2/3)m-\delta}\lambda \to \infty$ as $n \to \infty$ for some $\delta > 0$, then, as $n \to \infty$,

$$\int_\Omega (\hat{g} - g_0)^2 f_0 = o_p(n^{-1}\lambda^{-1/2m} + \lambda^2)$$

and this rate is exactly attained. Defining the Sobolev norm $\|g\|_S^2 = [g,g] + \int g^2 f_0$, provided $\lambda \to 0$ and $n\lambda^{(2m+1)/2m} \to \infty$ as $n \to \infty$, the estimator $\hat{g}$ is consistent for g_0 in Sobolev norm,

and, as $n \to \infty$,

$$\|\hat{g} - g_0\|_S^2 = o_p(n^{-1}\lambda^{-(2m+1)/2m}) + o_p(\lambda) .$$

The proofs of all the parts of the theorem are obtained by combining the relevant part of Theorem 6.1 with either Lemma 7.1 (for the L^2 and uniform consistency and rates) or Lemma 7.2 (for the Sobolev consistency and rates). In all cases it is the $\|g_1 - g_0\|$ part which dominates, the term $\|\hat{g} - g_1\|$ being negligible. The details are straightforward and are therefore omitted.

It is possible to investigate optimal rates of consistency and the corresponding rate of convergence of λ to zero. For mean square convergence, corresponding to convergence of the estimated density to the true density in the Kullback-Leibler information distance, the (exact) optimal rate of consistency is easily shown to be $O(n^{-4m/(4m+1)})$ attained when $\lambda \sim n^{-2m/(4m+1)}$. This rate of convergence near to $O(n^{-1})$ is of course a consequence of the strong smoothness conditions placed on g_0. The corresponding results for the other norms are left to the reader to investigate. It is interesting to note that the optimal rate for λ for good estimation in, for example, the Sobolev norm will <u>not</u> be the same as the optimal rate for mean square consistency. Thus one will not necessarily obtain good estimates of the derivatives of f_0 by seeking good estimates of f itself, a point relevant

to Silverman (1980) and the subsequent rejoinder of Good and Gaskins (1980).

The question of strong consistency, as considered for slightly different estimators by de Montricher (1981), is not considered in this paper, though it seems intuitively clear that analogous results to Theorem 8.1, possibly with slightly slower rates, should be provable by suitable techniques. The question of the asympotic normality of the estimates is considered in Section 9 below.

9. APPROXIMATION BY A GAUSSIAN PROCESS.

To the author's knowledge, roughness penalty density estimators are the only density estimators that have not been shown under suitaible conditions to be asymptotically normal. It turns out to be possible not only to show that the estimators discussed in this paper are pointwise asymptotically normal but also to give a rate of approximation to the estimators, suitably normalized, by a Gaussian process. An approximation of this kind is more in keeping with the modern theory of density estimators and opens the way to proving results such as those of Bickel and Rosenblatt (1973) and Silverman (1976) on the asymptotic behavior of certain functionals of the estimates. It shows that the joint distribution of the value of the estimator at several points is asymptotically multivariate normal and also gives a rate of convergence to this normal limit.

Before considering the approximation itself, it is convenient to consider some preliminaries. The notation and conventions of this section will be, except where otherwise stated, the same as in Sections 5 to 8 above. In particular all the conditions stated near the beginning of Section 5 will be assumed to hold. Defining the eigenfunctions ϕ_ν as in Section 5, define the function R_λ on $\Omega \times \Omega$ by

$$R_\lambda(s,t) = \sum_\nu \frac{\phi_\nu(s)\phi_\nu(t)}{(1 + \lambda\rho_\nu)} .$$

The function R_λ is the reproducing kernel with respect to the density f_0 corresponding to the inner product $\lambda[g_1,g_2] + \int g_1 g_2 f_0$ on S and is also the Green's function of a certain differential operator; for the connections, see a modern text on differential equations.

Define a function $m_\lambda(t)$ to be $Eg_1(t)$, so that the coefficients of m_λ satisfy

$$m_{\lambda\nu} = g_{0\nu}(1 + \lambda\rho_\nu)^{-1} .$$

It follows that

$$m_\lambda(s) = \int_\Omega R_\lambda(s,t) g_0(t) f(t) dt \qquad (9.1)$$

so that m_λ can be seen to be, in a certain sense, a smoothed version of g_0. Define the function r_λ by

$$r_\lambda(s,t) = \sum_{\nu=1}^{\infty} \frac{\phi_\nu(s)\phi_\nu(t)}{(1 + \lambda\rho_\nu)^2} ;$$

it can be shown easily that

$$r_\lambda(s,t) = \int_\Omega R_\lambda(s,u)R_\lambda(t,u)f_0(u)du - 1$$

and again $1 + r_\lambda$ is the Green's function of a certain differential operator. In addition, by methods similar to those used in the proof of Theorem 6.1, we have, for all s,

$$r_\lambda(s,s) \sim \lambda^{-1/2m} \quad \text{as} \quad \lambda \to 0 . \tag{9.2}$$

It is now possible to state and prove the main result of this section.

Theorem 9.1. Suppose that the conditions of Theorem 8.1 hold. For each n, on a suitable probability space there exists a Gaussian process $\gamma_\lambda(s)$ with mean zero and covariance function $r_\lambda(s,t)$ such that

$$g(s) = m_\lambda(s) + n^{-1/2}\gamma_\lambda(s) + \mathrm{err}_{n,\lambda}(s)$$

where the functions m_λ and r_λ are as defined above and, given $\delta > 0$, the approximation error $\mathrm{err}_{n,\lambda}$ is

$$o_p\{\lambda^{-\delta}(n^{-1}\lambda^{-1/m}\log n + \lambda^{-(4m-1)/2m})\}$$

uniformly over s in Ω as $\lambda \to 0$ and $n \to \infty$.

Proof.

The result is a consequence of Lemma 5.5 on the approximation of the β_ν by normal random variables. Note that the distribution of γ_λ does not depend on n, and so, as in Silverman (1976) and (1978a), theorems about its behavior as $\lambda \to 0$ can be combined

with Theorem 9.1 to provide results for g under transparent conditions connecting λ and n.

Define the β_ν as in Lemma 5.5 and define γ_λ by

$$\gamma_\lambda(s) = \sum_{\nu=1}^{\infty} \frac{\beta_\nu \phi_\nu(s)}{(1 + \lambda\rho_\nu)}$$

It is easily verified that γ_λ is a well defined random element of S and is a Gaussian process with $E\gamma_\lambda(s) = 0$ and $\text{cov}\{\gamma_\lambda(s), \gamma_\lambda(t)\} = r_\lambda(s,t)$. Using (6.3) it follows that the error process will satisfy

$$\text{err}_{n,\lambda}(s) = \sum_{\nu=1}^{\infty} \frac{(\beta_\nu - n^{-1/2}\tilde{\beta}_\nu)\phi_\nu(s)}{(1 + \lambda\rho_\nu)} + g(s) - g_1(s) \quad (9.3)$$

Using Lemma 5.2, the supremum over s of the sum in (9.3) is, given $\varepsilon > 0$, dominated by a constant multiple of

$$\{\sum \nu^{1+\varepsilon}(\beta_\nu - n^{-1/2}\tilde{\beta}_\nu)^2(1 + \lambda\rho_\nu)^{-2}\}^{1/2} \quad (9.4)$$

Now substitute the bound of Lemma 5.5 on $|\beta_\nu - n^{-1/2}\tilde{\beta}_\nu|$ to show that (9.4) is, with probability 1,

$$\{\sum \nu^{3+\varepsilon}(1 + \lambda\rho_\nu)^{-2}\}^{1/2} O(n^{-1}\log n) = O(\lambda^{-(4+\varepsilon)/4m} n^{-1}\log n)$$

by Lemma 5.3. Substituting this bound and (7.11) into (9.3) completes the proof of Theorem 9.1.

It is easy to use (9.2) and Theorem 9.1 to construct conditions under which $r_\lambda(t,t)^{-1/2}\{\hat{g}(t) - m_\lambda(t)\}$ is asymptotically standard normal and this is left to the reader to investigate.

10. DISCUSSION AND ACKNOWLEDGMENTS.

There are of course numerous questions still unanswered about roughness penalty density estimates. Apart from those technical questions raised in the body of this paper there are several important points of a practical nature which have not been discussed. Some heuristic calculations done by the author and not included here suggest that the estimates of this paper may provide a solution to the problem of finding estimates which are automatically adaptive to the tails of the distribution; most existing methods either under or oversmooth the tails relative to the main part of the data. Another important problem is the design of efficient and well understood data-based methods for choosing the smoothing parameter, though it should be the case that techniques from other density estimation methods can be adapted for use here. Finally it is of course important to have good computer algorithms for finding the estimates!

The author gratefully acknowledges useful discussions with Dennis Cox, Vassilios Klonias, Tom Leonard, Finbarr O'Suileabhain, Grace Wahba and James Wendelberger.

REFERENCES

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. Ann. Statist. 1, 1071-1095.

Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. Biometrika, 58, 255-277.

Good, I. J. and Gaskins, R. A. (1980). Density estimation and bump hunting by the penalized likelihood method exemplified by scattering and meteorite data (with discussion and rejoinder). J. Amer. Statist. Assoc., 75, 42-73.

Klonias, V. K. (1981). Consistency of a nonparametric penalized likelihood estimator of the probability density function. Preprint, Johns Hopkins University, Baltimore, Maryland.

Komlos, J., Major, P. and Tusnady, G. (1975). An approximation of partial sums of independent random variables, and the sample distribution function. Z. Wahrsch. Verw. Gebiete, 32, 111-131.

Leonard, T. (1978). Density estimation, stochastic processes and prior information (with discussion). J. Roy. Statist. Soc. Ser. B, 40, 113-146.

Mardia, K. V. (1972). Statistics of Directional Data. Academic Press, London.

de Montricher, G. M. (1981). On the consistency of maximum penalized likelihood estimators. Reprint, Rice University, Texas.

de Montricher, G. M., Tapia, R. A. and Thompson, J. R. (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods. Ann. Statist. 3, 1329-1348.

Oden, J. T. and Reddy, J. N. (1976). An Introduction to the Mathematical Theory of Finite Elements. J. Wiley, New York.

Reiss, R. D. (1981). Consistency of maximum penalized likelihood density estimators based on initial estimators. Preprint, Universitat-Gesamthochschule-Siegen, Siegen, W. Germany.

Rice, J. and Rosenblatt, M. (1981). Integrated mean square error of a smoothing spline. J. Approx. Theory, to appear.

Riesz, F. and Nagy, B. Sz-., (1955). Functional Analysis (trans. L. F. Boron). Ungar, New York.

Silverman, B. W. (1976). On a Gaussian process related to multivariate probability density estimation. Math. Proc. Cambridge Philos. Soc. 80, 135-144.

Silverman, B. W. (1978a). Weak and strong uniform consistency of the kernal estimate of a density and its derivatives. Ann. Statist., 6, 177-184.

Silverman, B. W. (1978b). Density ratios, empirical likelihood and cot death. J. Roy. Statist. Soc. Ser. C., 27, 26-33.

Silverman, B. W. (1980). Comment on Good and Gaskins (1980). J. Amer. Statist. Assoc., 75, 67-68.

Tapia, R. A. and Thompson, J. R. (1978). Nonparametric Probability Density Estimation. Johns Hopkins University Press, Baltimore, U.S.A.

Wahba, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy. SIAM J. Numer. Anal., 14, 651-667.

BS/ed

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 2228	2. GOVT ACCESSION NO.: AD-A103 875
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method	5. TYPE OF REPORT & PERIOD COVERED: Summary Report - no specific reporting period
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): B. W. Silverman	8. CONTRACT OR GRANT NUMBER(s): DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, P.O. Box 12211, Research Triangle Park, North Carolina 27709	12. REPORT DATE: June 1981
	13. NUMBER OF PAGES: 41
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

probability density estimate, roughness penalty, penalized likelihood, smoothing, data analysis, reproducing kernel Hilbert space, Sobolev space, convex optimization, existence and uniqueness, rates, consistency, asymptotic normality, Gaussian process, strong approximation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A class of probability density estimates can be obtained by penalizing the likelihood by a functional which depends on the roughness of the logarithm of the density. The limiting case of the estimates as the amount of smoothing increasing has a natural form which makes the method attractive for data analysis and which provides a rationale for a particular choice of roughness

(continued)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT (continued)

penalty. The estimates are shown to be the solution of an unconstrained convex optimization problem, and mild natural conditions are given for them to exist. Rates of consistency in various norms and conditions for asymptotic normality and approximation by a Gaussian process are given, thus breaking new ground in the theory of maximum penalized likelihood density estimation.